

OSSパッケージを用いた システム開発における OSSライセンスの互換性の検証支援手法

オープンソースソフトウェア工学研究室

M2 福井 克法

OSSの再利用

- ソフトウェアの開発コストを削減するために、ソフトウェアの再利用が行われる
- 多くの高品質なOSSが開発、公開されているため、多くの成果物の再利用が可能
- パッケージ管理システムによりOSSが手軽に利用可能
 - RPM Package Manager, Node Package Managerなど



OSSライセンス

- OSSを利用，修正，頒布を行う際にはソフトウェアに適用されているライセンスに準拠する必要がある
 - 再配布の条件
 - 公開義務
 - 他のOSSライセンスとの互換性
- ライセンス違反によるリスク
 - 損害賠償
 - 製品の打ち切り
 - 企業イメージの低下

OSSライセンスの互換性

- 異なるソフトウェアを一つのソフトウェアとして結合する場合、ライセンス間で矛盾がないか注意する必要がある
- 互換性のないライセンスを含む場合、ライセンス違反となる

リリースするソフトウェア

開発した
ソフトウェア



GPLv2.0

再利用した
ソフトウェア



Apache License 2.0

GPL v2.0 : 第6条抜粋

「あなたは、受領者がここで認められた権利を行使することに関してこれ以上他のいかなる制限も課してはならない」

Apache License 2.0 : 第3条抜粋

「あなたが誰かに対し、成果物が特許侵害に当たるとして特許訴訟を起こした場合、あなたに付与された特許ライセンスは終了する」

→ Apache Licenseの第3条の制約はGPLv2.0には無いため、
両ライセンスの互換性はなし（この場合GPLの条項に違反）

OSSライセンスの組み合わせに関する先行研究

- ソフトウェアパッケージ
 - パッケージで宣言されたライセンスとパッケージを構成するソースコードのライセンスの互換性を検証[1][2]
 - 依存関係にあるパッケージのライセンスの互換性を検証[3]
- Android アプリ
 - アプリを構成するソースコードとソースコード中で使用されているライブラリの互換性について検証[4]

[1]D.M. German, Y. Manabe, and K. Inoue, “A sentence-matching method for automatic license identification of source code files,” Proc. 25th International Conference on Automated software engineering (ASE '10), pp.437–446, 2010.

[2]Y. Manabe, D.M. German, and K. Inoue, “Analyzing the relationship between the license of packages and their files in free and open source software,” Proc. 10th IFIP WG2.13 International Conference on Open Source Systems (OSS '14), pp.51–60, 2014.

[3]M. Kechagia, D. Spinellis, and S. Androutsellis-Theotokis, “Open source licensing across package dependencies,” Proc. 14th Panhellenic Conference on Informatics (PCI '10), pp.27– 32, 2010.

[4]O. Mlouki, F. Khomh, and G. Antoniol, “On the detection of licenses violations in the android ecosystem,” Proc. 23rd International Conference on Software Analysis, Evolution and Reengineering (SANER '16), pp.382–392, 2016.

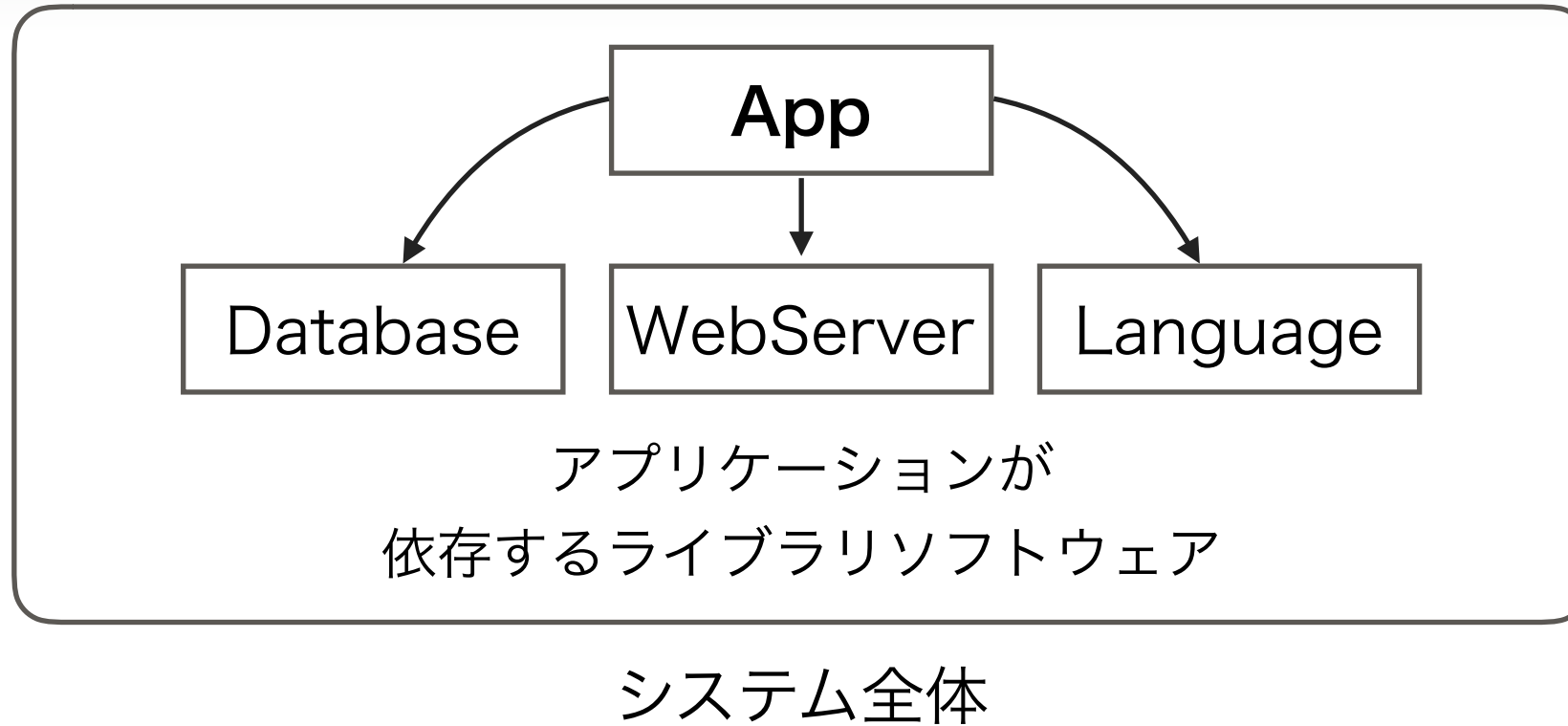
OSSライセンスの組み合わせに関する先行研究

- ソフトウェアパッケージ
 - パッケージで宣言されたライセンスとパッケージを構成するソースコードのライセンスの互換性を検証[1][2]
 - 依存関係にあるパッケージのライセンスの互換性を検証[3]
- Android アプリ
 - アプリを構成するソースコードとソースコード中で使用されているライブラリの互換性について検証[4]

ソフトウェアパッケージを組み合わせで構成されたシステム
におけるライセンスの互換性については検証されていない

[1]D
[2]Y
[3]M
[4]C

ソフトウェアパッケージを用いたシステム開発



- システム開発においては、複数のソフトウェアパッケージを組み合わせて1つのシステムを作成されることがある
- 仮想化技術（Dockerなど）の流行により、システム全体の配布が安易になっている

研究目的とアプローチ

- 研究目的

- 複数のソフトウェアパッケージを利用してシステム開発を行う際に、ソフトウェアパッケージ間のライセンスの互換性の検証を支援する

- アプローチ

1. 事前調査

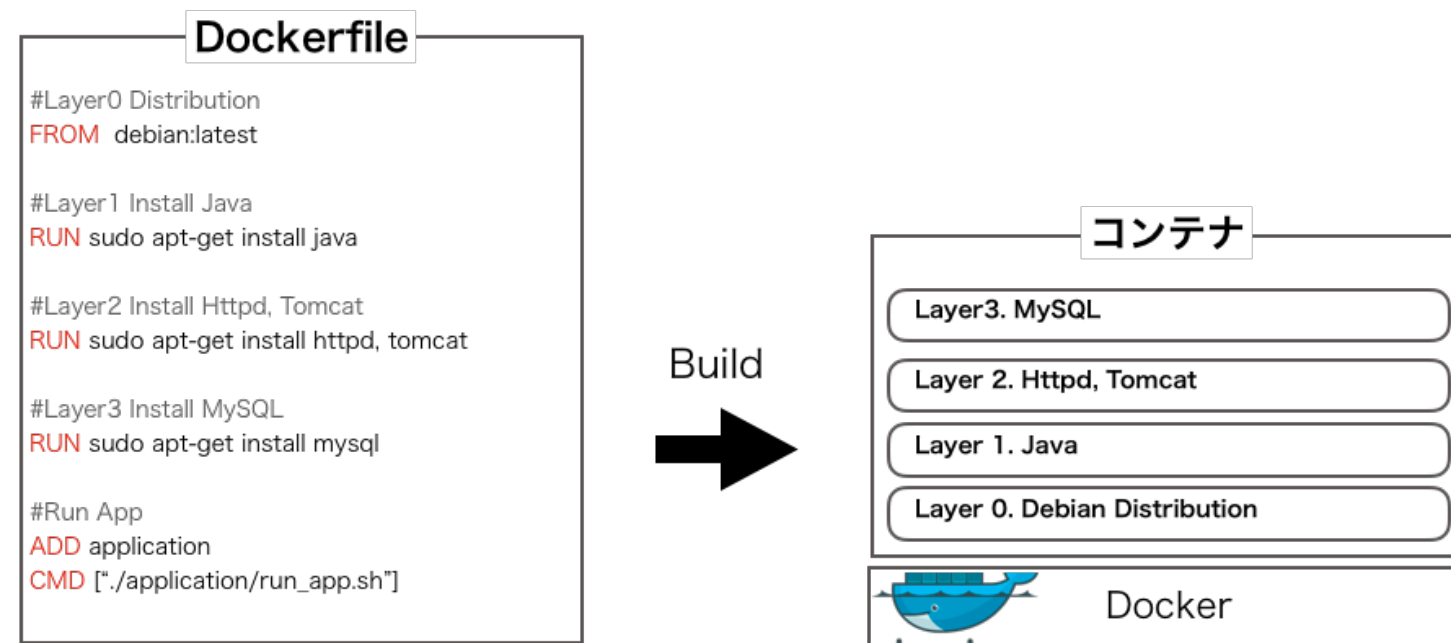
- ・ Dockerコンテナを対象として、ソフトウェアパッケージ間で互換性の問題が発生しているか調査

2. 互換性を持たないライセンスが適用されたソフトウェアパッケージが混入することを未然に防ぐ手法の構築

Docker

- Dockerコンテナ

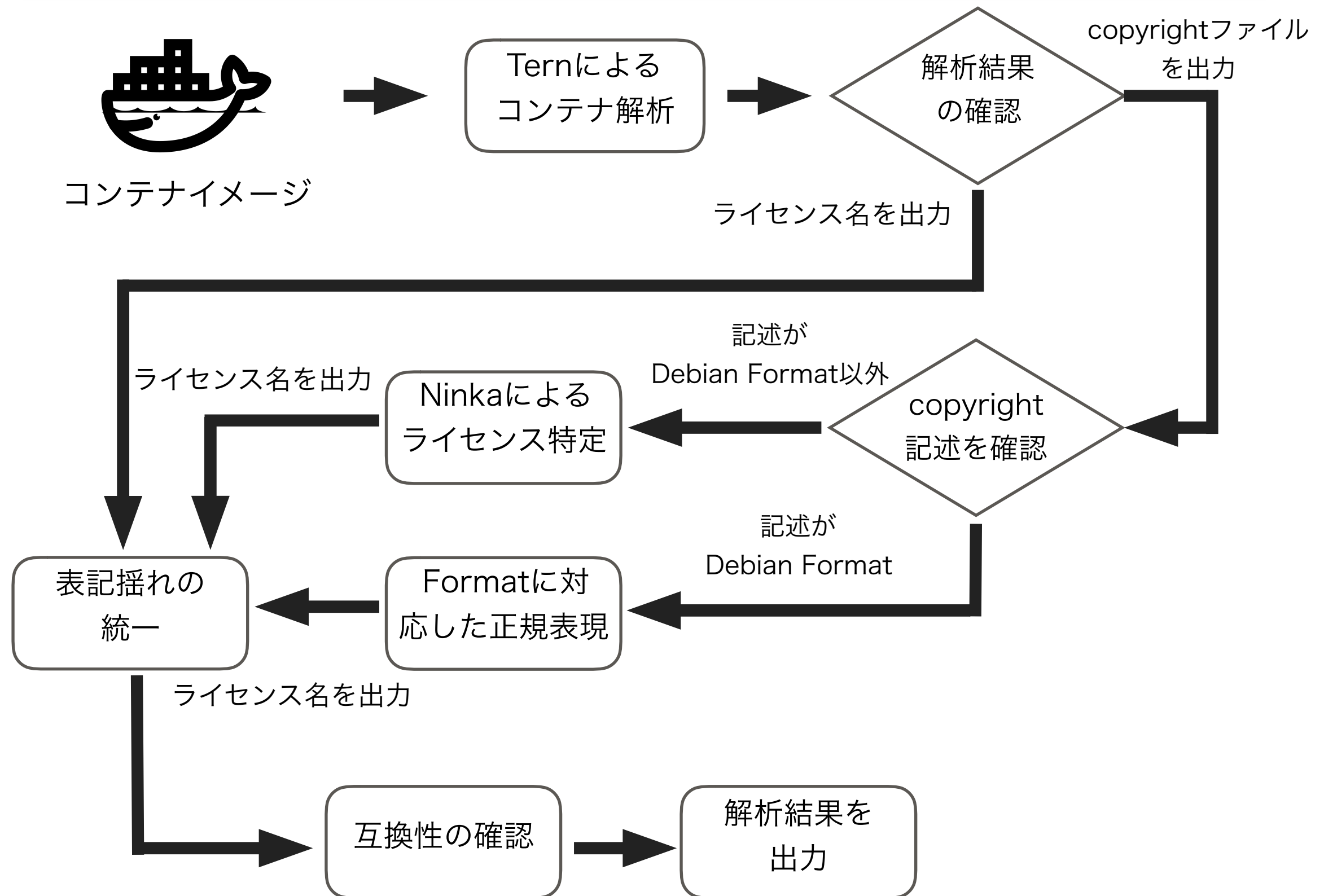
- アプリケーションと依存している
OSSパッケージをOSレベルで仮想化



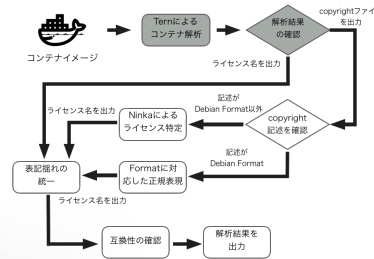
- Docker Hub（コンテナ共有サイト）

- 複数のOSSパッケージを組み合わせて開発されたシステム（コンテナ）の取得が可能

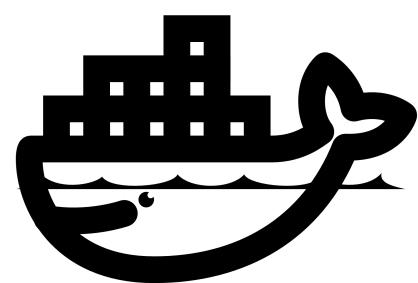
分析手法の概要



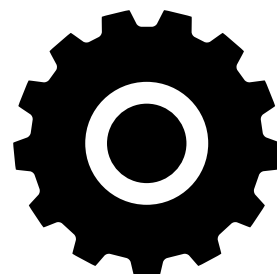
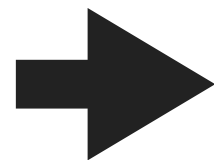
コンテナ解析ツール：Tern[5]



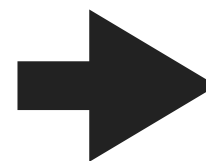
- コンテナを入力としてライセンスを特定する
- Debian, Ubuntu (APT管理システム) はCopyrightファイルの内容を出力する



解析対象
コンテナ



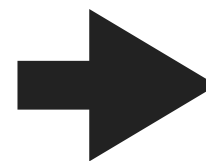
コンテナ解析ツール
Tern



CentOS
Alpineなど

差分情報	ソフトウェア	ライセンス
03901b4a2ea88ee	musl	MIT
03901b4a2ea88ee	busybox	GPL-2.0
03901b4a2ea88ee	alpine-keys	MIT
03901b4a2ea88ee	libcrypto1.1	OpenSSL
03901b4a2ea88ee	zlib	zlib
03901b4a2ea88ee	scanelf	GPL-2.0
03901b4a2ea88ee	libc-utils	BSD

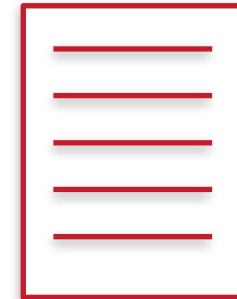
各レイヤーでインストールされた
ライセンス一覧



Debian
Ubuntu

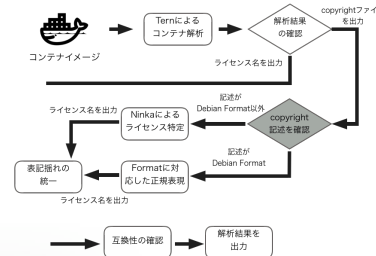


ライセンス文



その他の
format

Copyrightの出力例



- APT管理システムを用いている場合はCopyrightファイルの内容を出力
 - 目視により特定できたフォーマット

• ライセンス文

This package was written by Peter Tobias <tobias@et-inf.fho-emden.de>
on Thu, 16 Jan 1997 01:00:34 +0100.

License:

Copyright (C) 2009 Michael Meskes <meskes@debian.org>
Copyright (C) 2004-2005 Graham Wilson <graham@debian.org>
Copyright (C) 1997 Bernd Eckenfels
Copyright (C) 1997 Peter Tobias <tobias@et-inf.fho-emden.de>
Copyright (C) 1996 Free Software Foundation, Inc.

This program is free software; you can redistribute it and/or modify
it under the terms of the GNU General Public License as published by
the Free Software Foundation; version 2 of the License.

This program is distributed in the hope that it will be useful,
but WITHOUT ANY WARRANTY; without even the implied warranty of
MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See

the

GNU General Public License for more details.

You should have received a copy of the GNU General Public License
along with this program; if not, write to the Free Software
Foundation, Inc., 51 Franklin St, Fifth Floor, Boston, MA 02110-1301 USA

• Debian Format

Format: <https://www.debian.org/doc/packaging-manuals/copyright-format/1.0/>
Upstream-Name: grep
Upstream-Contact: bug-grep@gnu.org
Source: <https://savannah.gnu.org/projects/grep>
[git://git.savannah.gnu.org/grep.git](https://git.savannah.gnu.org/grep.git)

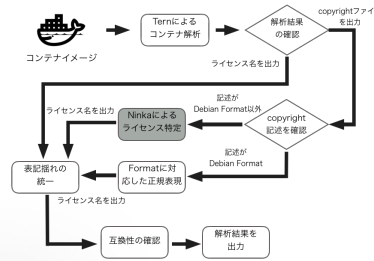
Files: *

Copyright: 1992, 1997-2002, 2004-2012 Free Software Foundation, Inc.
2004, Stepan Kasal <kasal@ucw.cz>
2007, Tony Abou-Assaleh <taa@acm.org>
2009-2012, Jim Meyering <jim@meyering.net> and Paolo Bonzini
<bonzini@gnu.org>
License: GPL-3+

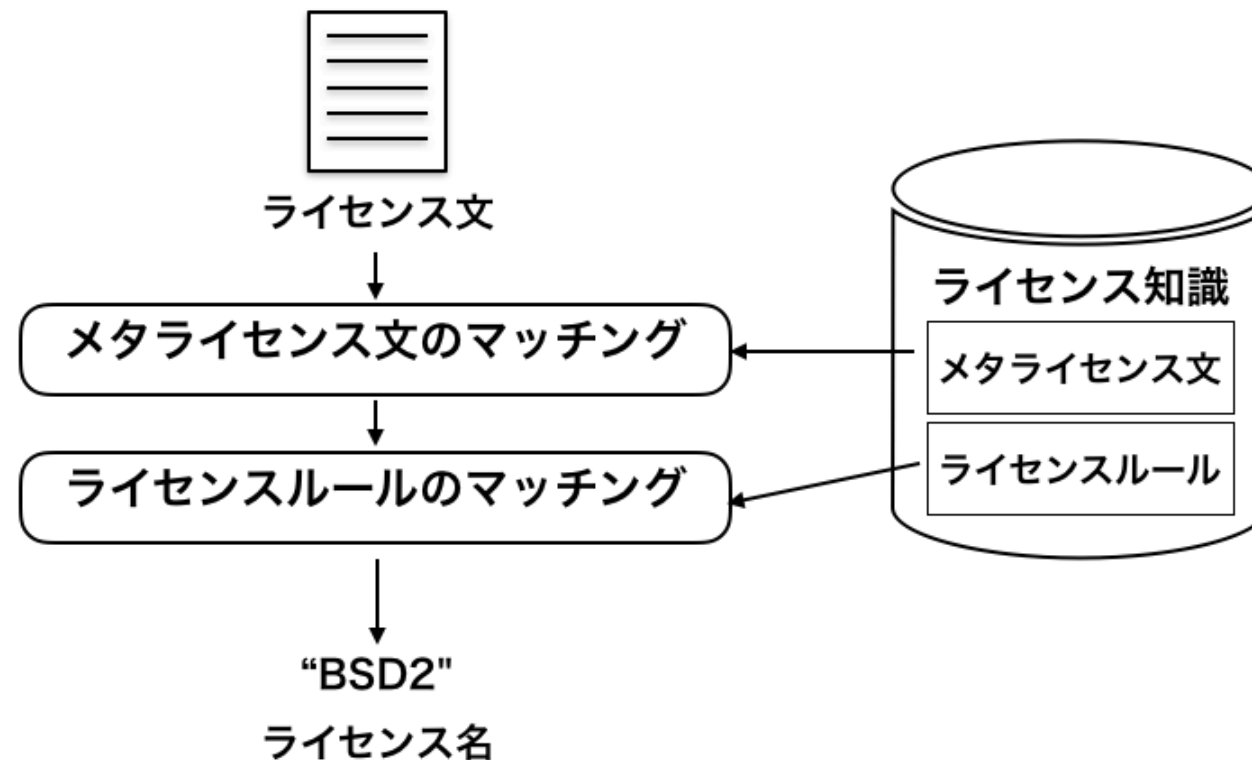
Files: debian/*

Copyright: 2005-2013 Anibal Monsalve Salazar <anibal@debian.org> and Santiago
Ruano Rincón <santiago@debian.org>
2003-2004 Ryan M. Golbeck <rmgolbeck@debian.org>
2003, Jeff Bailey <jbailey@nisa.net>
2003, Clint Adams <schizo@debian.org> Mon, 10 Mar 2003 02:10:32 -0500
2001 Robert van der Meulen <rvdm@debian.org>
1996-2000 Wichert Akkerman <wakkerma@debian.org>
License: GPL-3+

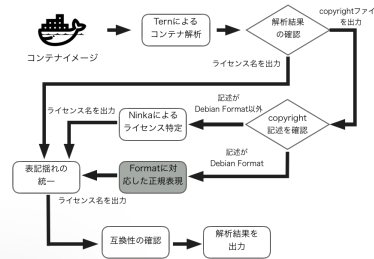
Ninka[6]



- Copyrightファイルに記述されるライセンス文の
ライセンス特定にはNinkaを用いる
 - 複数あるライセンス特定手法の中で、もっとも精度が高い
 - ・ 適合率 96.6%
 - 112個のライセンスに対応
(ライセンス知識を用いて特定)



Debian Formatのライセンス特定



- Copyrightファイルの記述がDebian Formatの場合は、正規表現を利用してライセンス名を特定する
 - 1.各ファイルのライセンスが書かれた記述をブロック単位に分割
 - 2.File : *を含むブロックのライセンスを正規表現で特定

Format: <https://www.debian.org/doc/packaging-manuals/copyright-format/1.0/>
Upstream-Name: grep
Upstream-Contact: bug-grep@gnu.org
Source: <https://savannah.gnu.org/projects/grep>
git://git.savannah.gnu.org/grep.git

Files: *

Copyright: 1992, 1997-2002, 2004-2012 Free Software Foundation, Inc.
2004, Stepan Kasal <kasal@ucw.cz>
2007, Tony Abou-Assaleh <taa@acm.org>
2009-2012, Jim Meyering <jim@meyering.net> and Paolo Bonzini <bonzini@gnu.org>
License: GPL-3+

Files: debian/*

Copyright: 2005-2013 Anibal Monsalve Salazar <anibal@debian.org> and Santiago Ruano Rincón <santiago@debian.org>
2003-2004 Ryan M. Golbeck <rmgolbeck@debian.org>
2003, Jeff Bailey <jbailey@nisa.net>
2003, Clint Adams <schizo@debian.org> Mon, 10 Mar 2003 02:10:32 -0500
2001 Robert van der Meulen <rvdm@debian.org>

1.

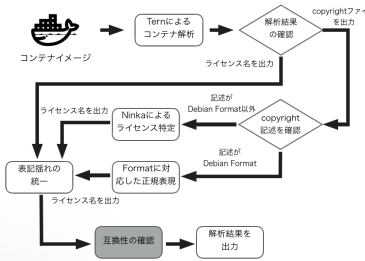
Files: *

Copyright: 1992, 1997-2002, 2004-2012 Free Software Foundation, Inc.
2004, Stepan Kasal <kasal@ucw.cz>
2007, Tony Abou-Assaleh <taa@acm.org>
2009-2012, Jim Meyering <jim@meyering.net> and Paolo Bonzini <bonzini@gnu.org>
License: GPL-3+

2.正規表現により
ライセンス名を抽出

License: GPL-3+

ライセンスの互換性の確認



- コンテナに含まれるライセンスから非互換の組み合わせを作成
 - GPLの公式サイト[7], 既存研究[8]を参考に作成
 - ・ ライセンス名の表記ゆれは目視により統一

ライセンスの組み合わせ	互換性の有無
GPLv2+_Apache1	×
GPLv2+_BSD4	×
GPLv2+_MPLv1_1	×
GPLv2+_CDDL-1.0	×
GPLv2+_EPL-1.0	×
GPLv2+_LPPL	×
GPLv2+_OpenSSL	×
GPLv2_GPL3+	×
GPLv2_LGPL3+	×
GPLv2_Apache1	×

GPLv2_Apache2.0	×
GPLv2_BSD4	×
GPLv2_MPLv1_1	×
GPLv2_CDDL-1.0	×
GPLv2_EPL-1.0	×
GPLv2_LPPL	×
GPLv2_OpenSSL	×
GPLv3+_Apache1	×
GPLv3+_BSD4	×
GPLv3+_MPLv1_1	×
GPLv3+_CDDL-1.0	×
GPLv3+_EPL-1.0	×
GPLv3+_LPPL	×
GPLv3+_OpenSSL	×

[7]<https://www.gnu.org/licenses/license-list.html>

[8]D. M. German and J. M. Gonzalez-Barahona. An empirical study of the reuse of software licensed under the GNU General Public License. In Proc. of the International Open Source Systems Conference (OSS'09), Springer, 2009, 185-198.

事前調査で用いるデータセット

関連研究[9]のデータより利用履歴のあるコンテナのリストを抽出

	コンテナ数
公式コンテナ	156
非公式コンテナ	620
合計	776

- データの前処理（対象データ）
 - 対象コンテナ
 - ・ Docker Hubからダウンロードが可能なコンテナのみ
 - ・ 公式のDistributionコンテナ以外のコンテナのみ
 - ・ Ternで解析可能なコンテナのみ
 - 対象ソフトウェアパッケージ
 - ・ ライセンス特定で、ソフトウェアパッケージに対して一意にライセンスが特定できたもののみ

事前調査の結果

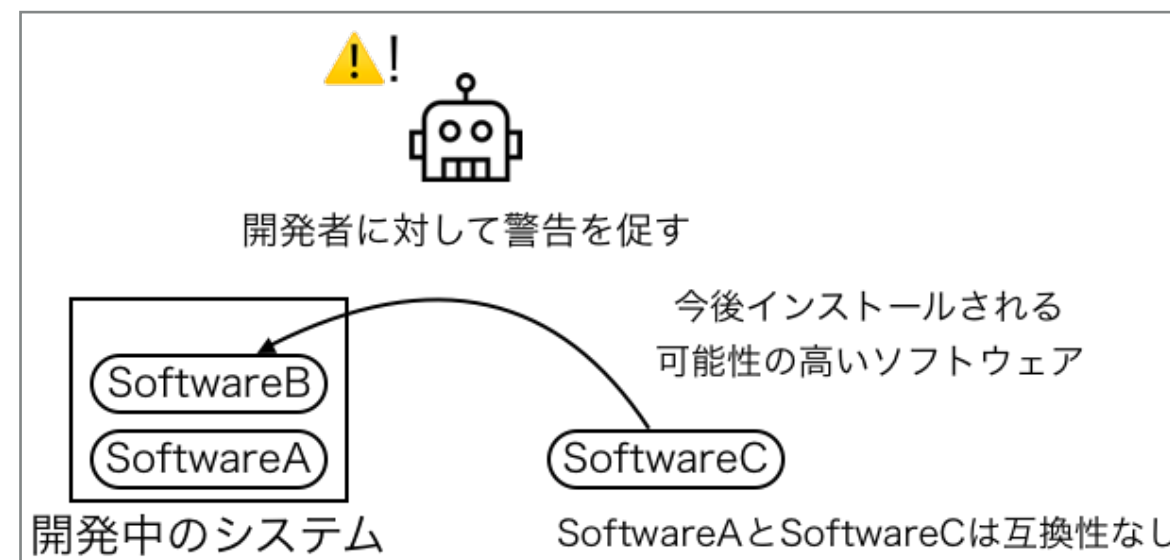
非互換の組み合わせを含むかの調査結果

	非互換の組み合わせを含む	非互換の組み合わせを含まない	合計
公式コンテナ	119	37	156
非公式コンテナ	338	232	620
合計	457	319	776

- 約半数のコンテナで互換性を持たないライセンスが適用されたソフトウェアパッケージの組み合わせを含んでいる
- ➡ 開発したシステムを安全に使うためには、未然に互換性を持たない組み合わせを含むことを防止することが必要

非互換の組み合わせを含むことを予測する手法

- 予測手法構築の目的
 - 互換性を持たないライセンスが適用されているソフトウェアパッケージを含むことを，システム開発の初期段階で未然に防ぐ



手法のイメージ

- アプローチ
 - 相関ルールマイニングを用いた予測手法の構築を行う
 - ・ 頻繁に組み合わせるソフトウェアパッケージのパターンを取り出す

相関ルールマイニング

$$X \Rightarrow Y$$

X：前提部

Y：結論部

Xという条件が満たされる場合に同時にYという条件も満たされるルールを抽出する手法

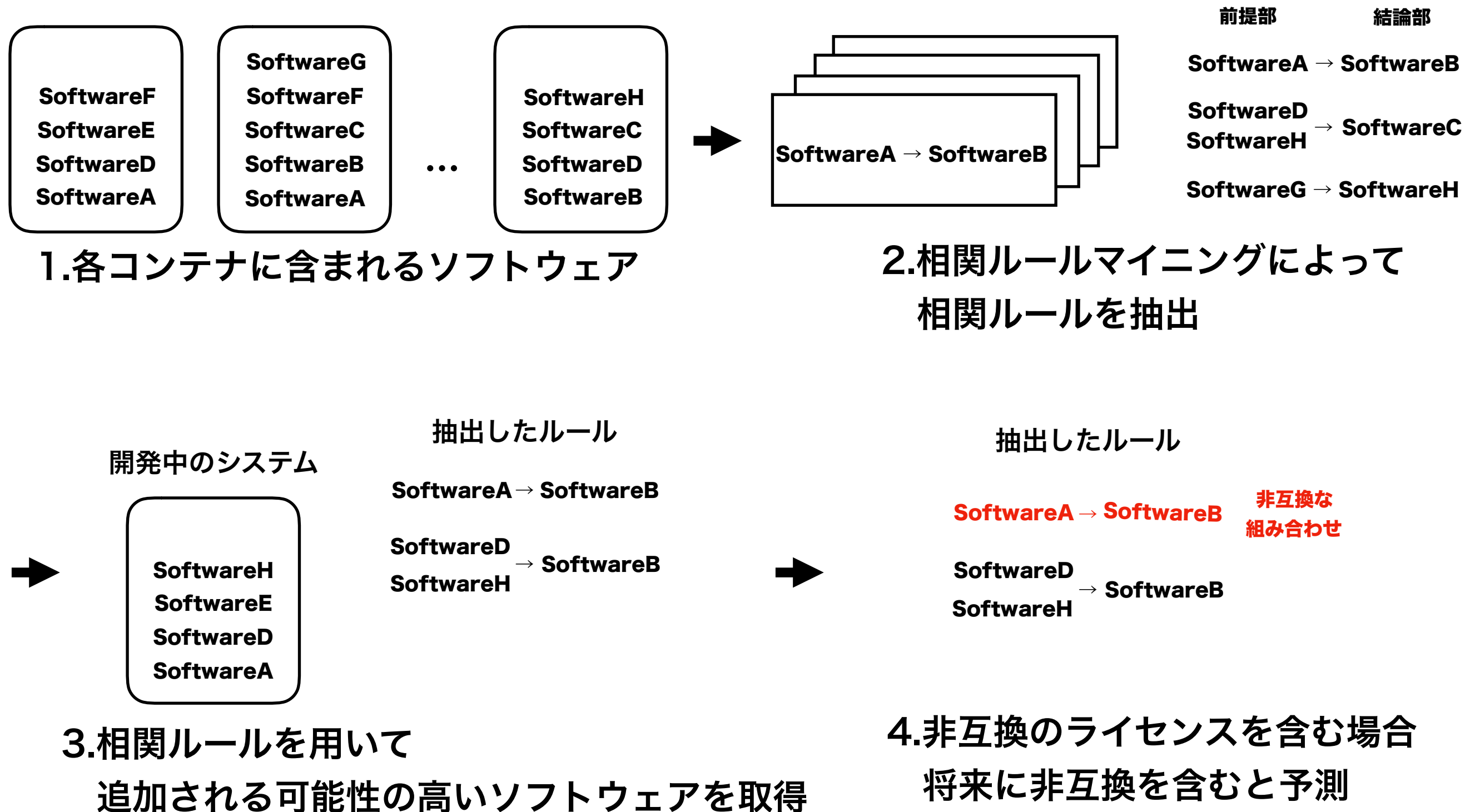
例：{MySQL, PHP} $\Rightarrow$ {Apache}

- support, confidenceの閾値を設定しルールを抽出

$$support(X, Y) = \frac{\text{条件Xを満たすコンテナ数}}{\text{全コンテナ数}}$$

$$confidence(X, Y) = \frac{\text{条件XとYを同時に満たすコンテナ数}}{\text{条件Xを満たすコンテナ数}} = \frac{support(X \cup Y)}{support(X)}$$

相関ルールマイニングを用いた予測手法



評価に用いるデータセット

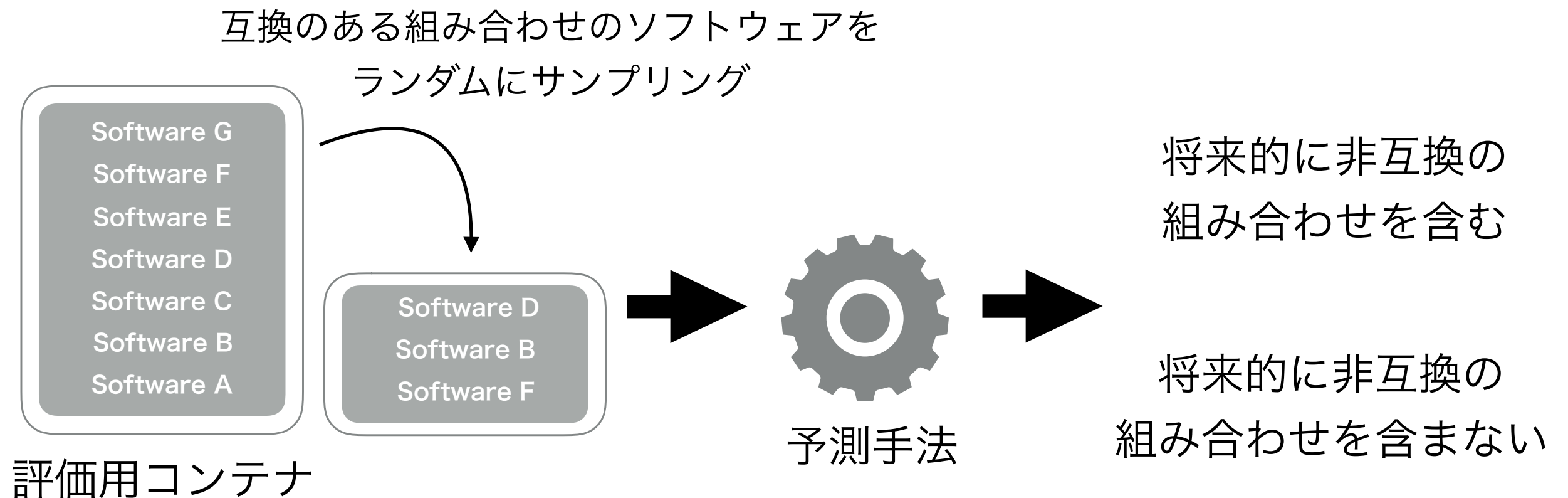
予測手法に用いるデータセット

非互換の組み合わせを含む	非互換の組み合わせを含まない	合計
457(77%)	141(23%)	598

- 分析で得られた結果を元にデータセットを構築
- データのフィルタリング
 - Distributionのみで構成されているコンテナは除去

実験方法

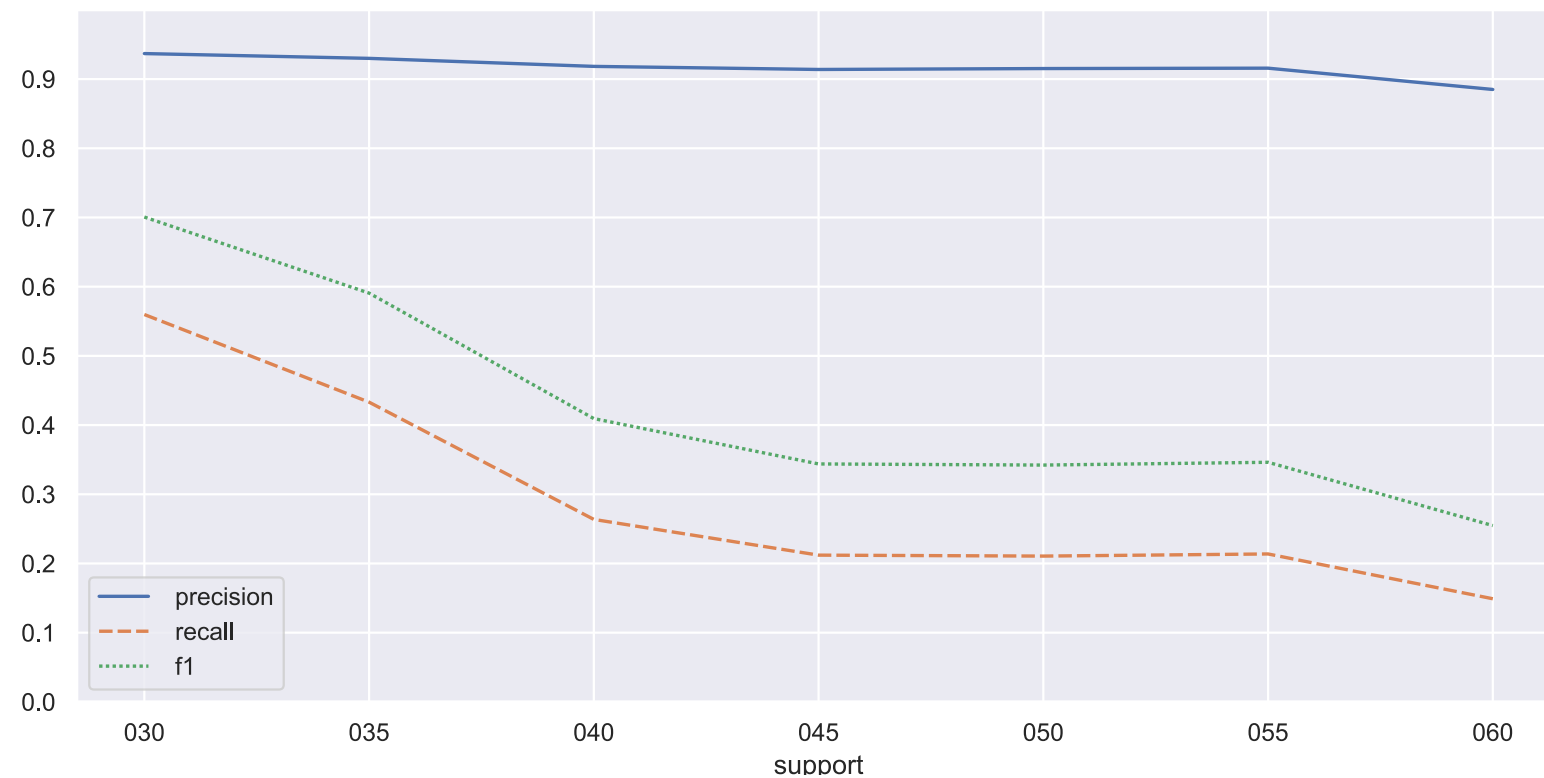
- すでにコンテナから一部のソフトウェアを取り出し、一部のソフトウェアからどの程度、将来の非互換性を予測できるか検証



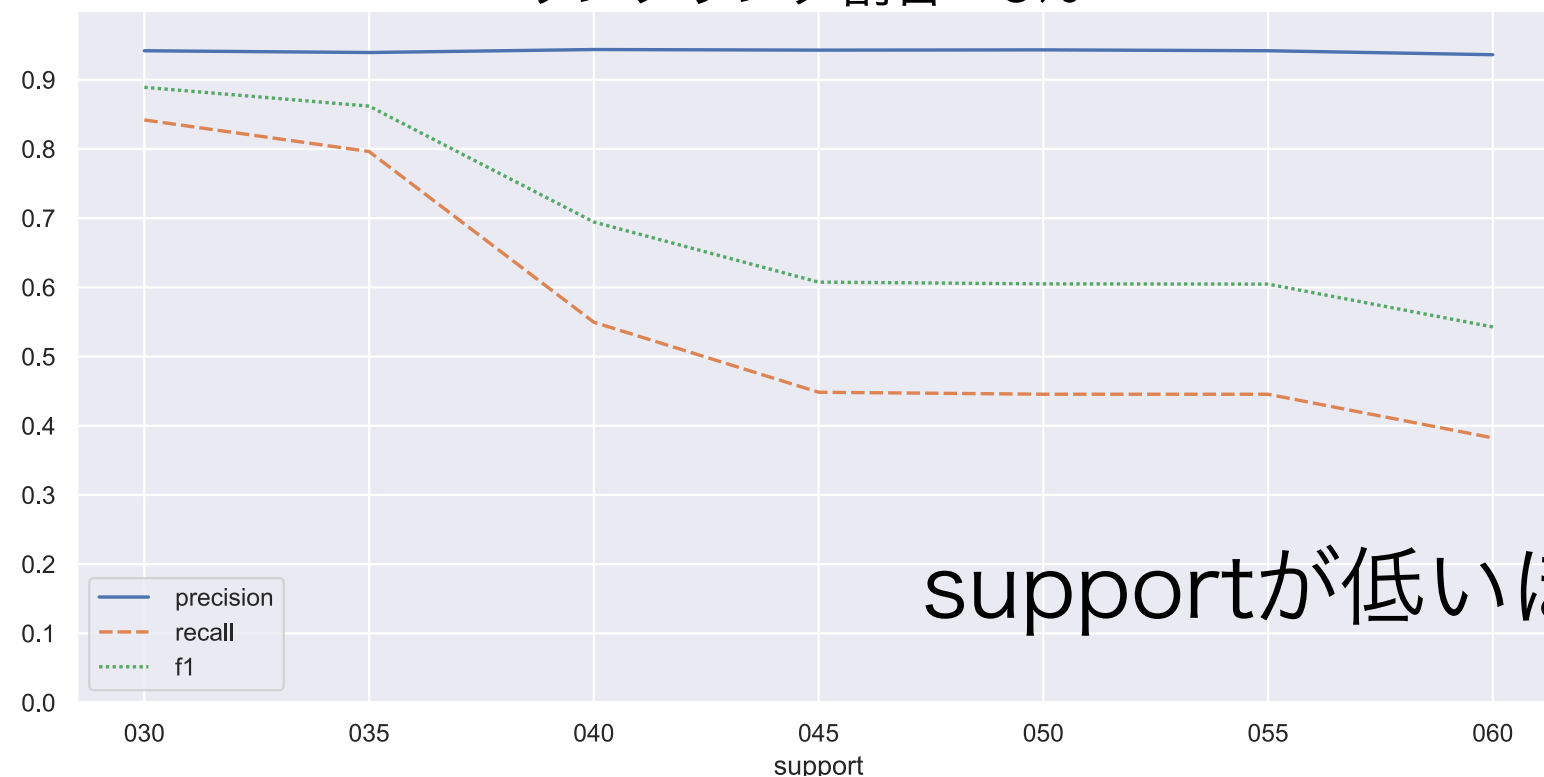
評価指標

- 実験方法
 - Leave-one-out交差検証（100回）
 - ・ 1つのコンテナを検証用として，その他のコンテナでルールを作成し評価を行う．また，同様の手順を全コンテナパターン実施
- 評価指標
 - Precision, Recall, F1
- パラメータの変更方法
 1. Confidence : 0.5, Support:0.3 ~ 0.6で変化（0.05）
 2. 1より最もF1の高いSupportに固定，
Confidence:0.05 ~ 0.95で変化（0.05）

結果1 (Confidence : 0.5に固定)



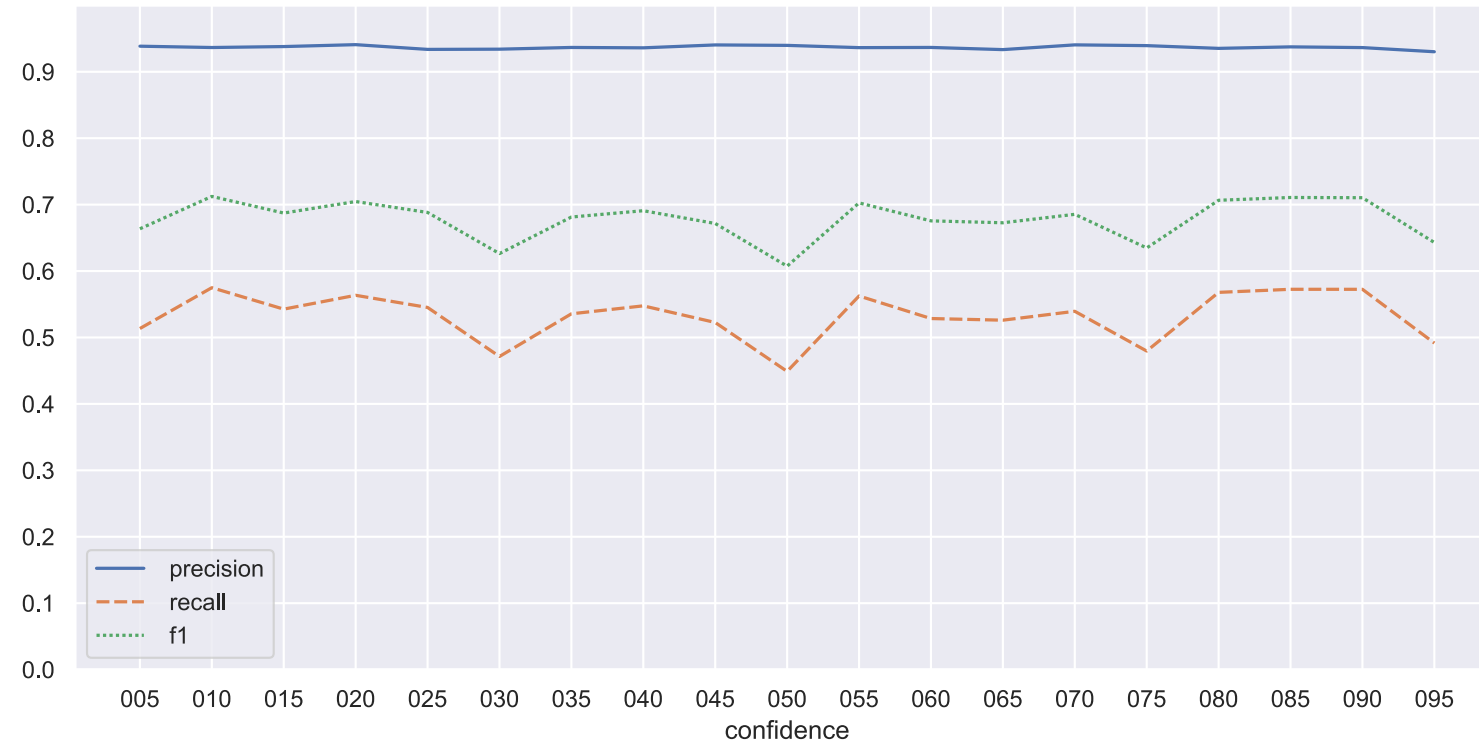
サンプリング割合：5%



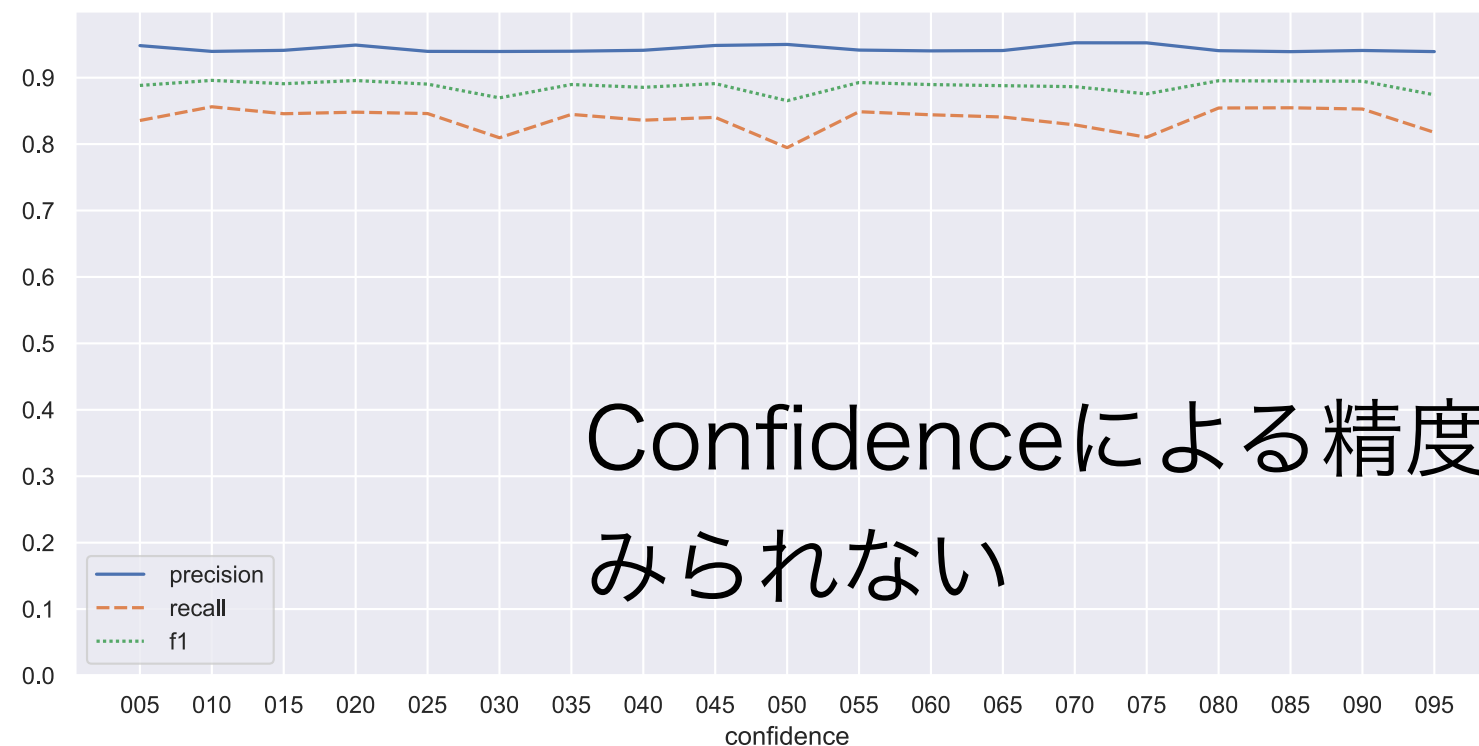
supportが低いほど精度が高い

サンプリング割合：30%

結果2(Support : 0.3に固定)



サンプリング割合：5%

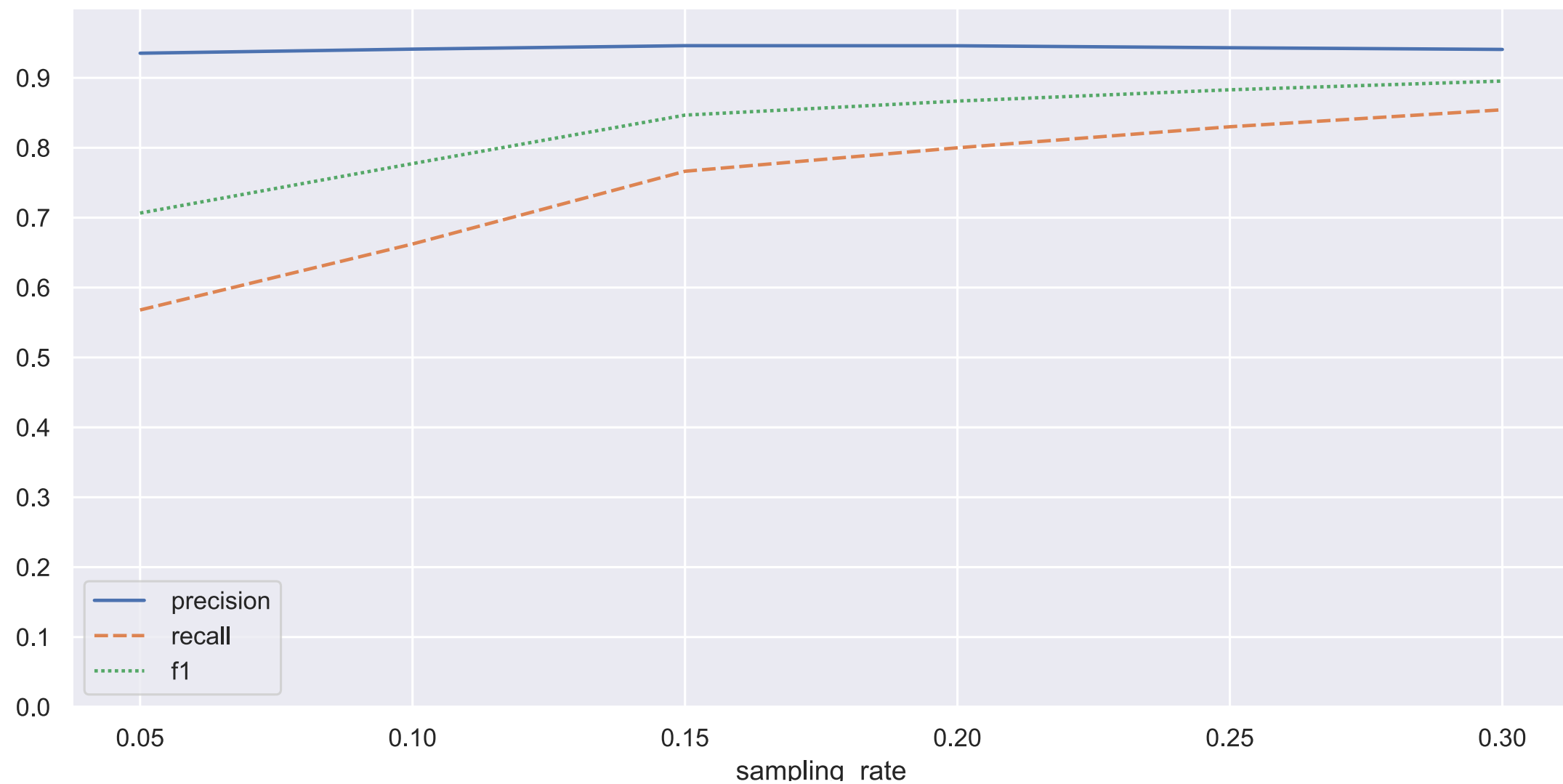


Confidenceによる精度の大きな変化は
みられない

サンプリング割合：30%

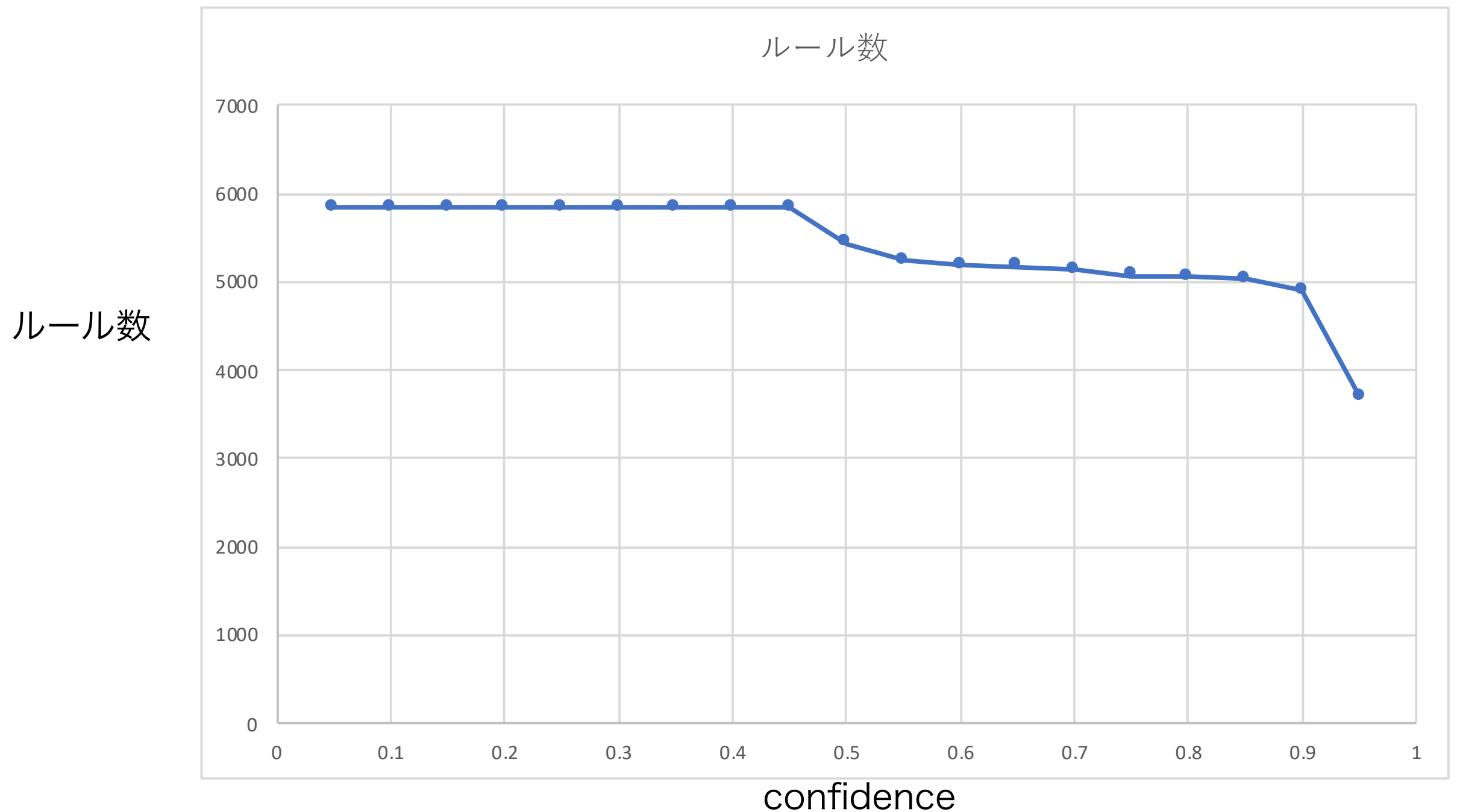
結果3 (サンプル数による精度の変位)

- 支持度：0.3 確信度：0.8



Support 0.30のときのルール数の変化

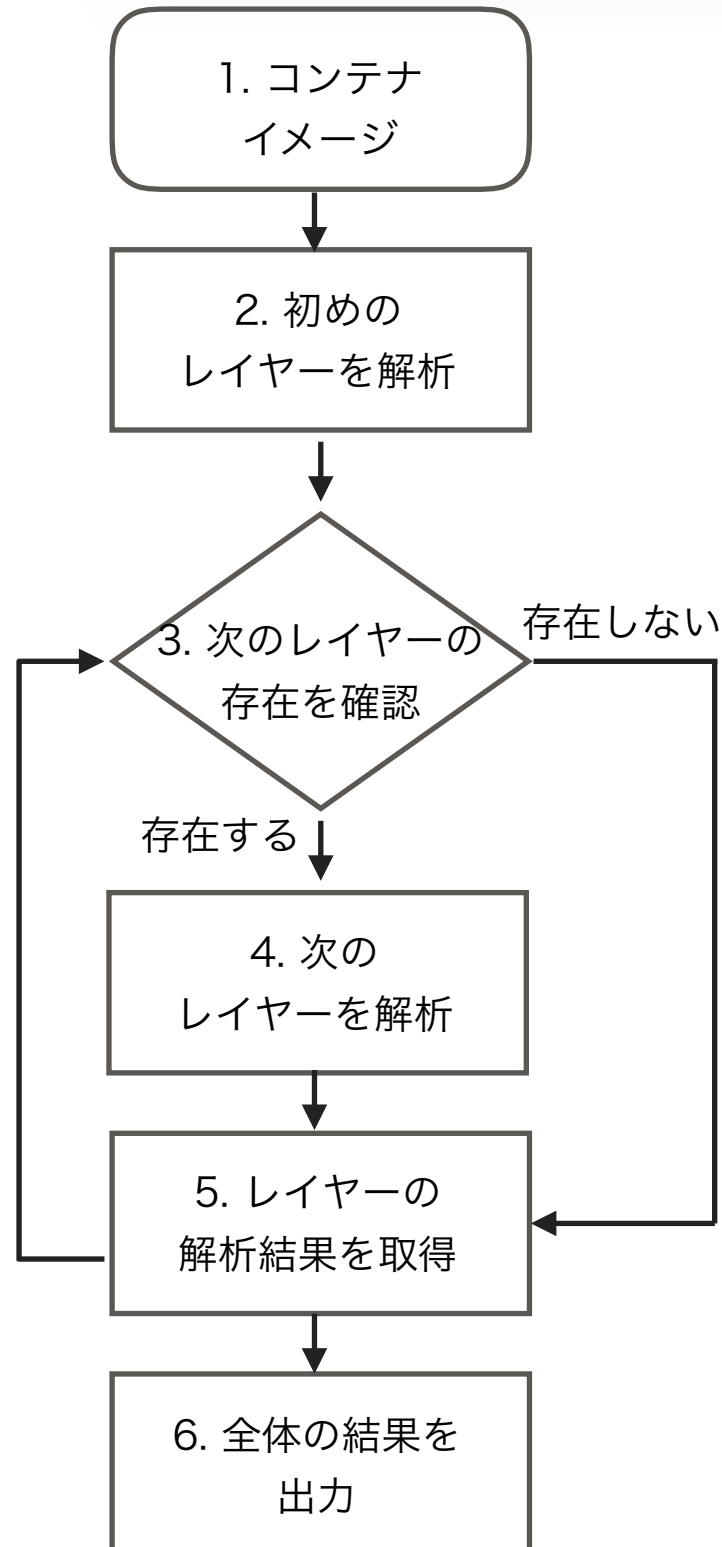
- Support : 0.3としてConfidenceを変化させたときのルール数の変化



まとめ

- まとめ
 - Leave one outによる実験終了
- 今後の予定
 - 評価時のデータの分割方法を変えて, supportをさらに下げれるか検証

Ternの詳細



1. コンテナの構成ファイル（tar）を展開
1番目のレイヤーをchrootコマンドでホストマシンにマウント
2. os-releaseファイルからOSを特定
パッケージ管理ソフトウェアで
メタ情報からインストールソフトウェアを特定
3. 次のレイヤーが存在するか確認
4. 次のレイヤー構成フォルダを結合し，ホストマシンにマウント
手順2と同様にインストールソフトウェアを特定
5. 各レイヤーのインストールソフトウェア情報を取得
6. 各レイヤーごとに全体の結果を出力

Ninkaの詳細

一般的なBSD2のライセンス記述

Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

1. Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.
2. Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.

BSD2のメタライセンス文

ルール名 : BSDpre

正規表現 : Redistribution(s)? and use in source and binary[...]

ルール名 : BSDcondSource

正規表現 : Redistribution(s)? of source code[...]

ルール名 : BSDcondBinary

正規表現 : Redistribution(s)? in binary form[...]

- メタライセンス文
 - ライセンス文に対応した正規表現
- BSD2のライセンスルール
 - BSD2:BSDpre,BSDcondSource,BSDcondBinary,BSDasIs,BSDWarr



Ternによる
コンテナ解析

解析結果
の確認

ライセンス名を出力

ライセンス名を出力

Ninkaによる
ライセンス特定

記述が
Debian Format以外

copyright
記述を確認

記述が
Debian Format

Formatに対
応した正規表現

表記揺れの
統一

ライセンス名を出力

互換性の確認

解析結果を
出力

copyrightファイル
を出力

